NSTL 集成利用第三方来源元数据的实践与探索*

于倩倩 张建勇
(中国科学院文献情报中心 北京 100190)

摘要: 【目的】将 WOS、Scopus 等第三方来源元数据应用到 NSTL 加工系统中。【应用背景】根据 NSTL 发展规划，需要从单纯自加工扩展到加工以及协商获取、购买第三方元数据等多渠道建设元数据方式。【方法】以 NSTL 加工规范为基础，实现与 WOS、Scopus 元数据的映射，分析第三方元数据特点对 NSTL 加工规范进行局部修订并映射，根据映射结果，将第三方元数据以 NSTL 加工规范格式输出并集成到 NSTL 加工系统中。【结果】实现第三方来源元数据快速、高效、低成本地集成整合到 NSTL 加工系统。【结论】WOS 元数据在 NSTL 加工系统中的应用，可以提高 NSTL 文献数据加工速度。有针对性地对现有元数据加工规范进行修订，为后续增加其他第三方资源构建了拓展框架。

关键词: Web of Science Scopus NSTL 元数据映射
分类号: G250.7

1 引 言

国家科技图书文献中心(NSTL)“十三五”发展规划提出，要优化国家科技文献资源保障体系，拓展元数据资源采集方式。为此，要整合、集成和利用第三方来源元数据，从单纯自己加工扩展到加工以及与国内外出版商、相关信息机构等第三方协商获取、交换、赠与、呈缴和购买等多渠道建设元数据资源方式。因此，需要在 NSTL 采用的文献资源元数据加工规范[1]基础上，深入分析其他来源元数据的类型特点和建设需求，建立健全 NSTL 元数据规范，以便更加有效地集成利用第三方来源元数据。

目前，不同的文献数据库，元数据的内容和描述方式存在差异，这对集成和利用第三方资源产生障碍。元数据格式的多样性与 NSTL 加工规范需求接口单一性之间的冲突，使得第三方来源元数据与 NSTL 文献资源元数据之间的互操作成为必然[2-4]。明确外部来源元数据的内容和组织方式，制订相关规则实现第三方来源元数据与 NSTL 文献资源元数据的映射，并将外部来源元数据资源以 NSTL 元数据格式输出，是 NSTL 集成利用第三方数据库数据的可操作方式之一。

Web of Science(简称 WOS)数据库[5]、Scopus 数据库[6]是国际知名的数据库，在提供文献信息服务方面与 NSTL 具有相同之处。本文在分析 WOS 元数据规范、Scopus 元数据规范和 NSTL 采用的文献资源加工规范基础上，结合相关实践，以期刊论文为例，对三者的元数据映射内容、映射效果、元数据描述方式进行比较，并提出映射及利用第三方元数据过程中需要注意的问题，以期为相关文献信息系统的元数据建设和利用已有第三方来源元数据资源提供借鉴。

2 期刊论文元数据结构

根据 DC 元数据设计的模块化原则[7]，并结合分析 WOS、Scopus、NSTL 三个文献数据库的元数据内容，期刊论文元数据可以分为论文元数据、作者

通讯作者：于倩倩，ORCID：0000-0001-8777-1171，E-mail：yuqianqian@mail.las.ac.cn。
*本文系 NSTL 支持项目“数据加工流程调整和加工系统改造”(项目编号:2014XM076)的研究成果之一。

元数据、作者机构元数据、期刊元数据、会议元数据、基金元数据、参考文献元数据、施引文献元数据等。按照实体分析法，期刊论文实体间的关系如图 1 所示，一篇期刊论文可能由一个或多个作者撰写，一个作者属于一个或多个机构，论文发表在期刊上，可能来自于某个会议，也可能挂靠某个基金，可能具有一篇或多篇参考文献，也可能被一篇或多篇文献引用等。

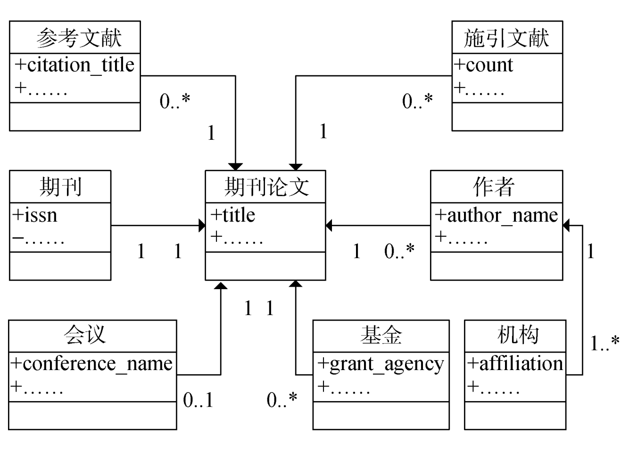


图 1　期刊论文实体关系

三个数据库包含的期刊论文元数据类型有所不同，如表 1 所示，可以看出，WOS、Scopus 对 8 类元数据均有描述，NSTL 缺乏对会议、基金和施引文献元数据的描述。原因主要是：WOS、Scopus 使用一套元数据 Schema 描述多种文献类型如期刊论文、会议论文、图书、专利等。因此，如果期刊论文中涉及到会议、基金信息，会出现相关描述，NSTL 以文献类型为基础划分元数据 Schema，会议元数据包含在会议论文 Schema 中；NSTL 加工规范没有对基金数据、施引文献数据的描述。

表 1　WOS、Scopus、NSTL 的期刊论文元数据

元数据类型	论文	作者	作者机构	期刊	会议	基金	参考文献	施引文献
WOS	√	√	√	√	√	√	√	√
Scopus	√	√	√	√	√	√	√	√
NSTL	√	√	√	√			√	

3　元数据映射与比较

以 NSTL 期刊论文元数据(部分字段是必备(Required)字段，以 R 表示)为基础，对比 WOS、Scopus 在论文元数据、作者/机构元数据、期刊元数据、参考文献元数据中相同字段的描述内容和方式，并分析不同文献数据库元数据描述的特点，以期取长补短，改善 NSTL 文献资源加工规范元数据的完整性和兼容性，更好地适应和支撑对各类第三方来源元数据的集成整合。

3.1　论文元数据的映射比较

NSTL 论文描述信息是期刊论文描述元数据规范的主体部分，描述的内容包括论文题名、关键词、文摘和分类信息等几个部分。WOS、Scopus 中与 NSTL 论文描述信息等同的字段来源于不同的元数据模块。例如, WOS 中题名、文献类型信息来源于论文元数据，起页、止页、总页数来源于期刊元数据; Scopus 中题名、摘要、文献类型信息来源于论文元数据，起页、止页、总页数来源于期刊元数据，参考文献总数来源于参考文献元数据等。WOS、Scopus 与 NSTL 论文元数据映射如表 2 所示。

从表 2 可以看出，在 22 个 NSTL 论文元数据字段中, WOS 有 12 个字段实现映射, Scopus 有 16 个字段实现映射，不同的第三方来源元数据与 NSTL 元数据映射数量不同。在未映射的字段中，包含了必备字段 paper_id 和 local_doi，必须对这两个必备字段进行处理才能将映射后的外部数据源数据以 NSTL Schema 格式输出，例如将必备字段输出为空标签。

在实现映射的字段中，同一字段在不同数据库中取值、字段可重复性不同，也对外部数据源数据输出为 NSTL Schema 格式造成影响。例如, NSTL 的 type、WOS 的 doctype、Scopus 的 citation-type，虽然都是描述文献的类型，但三者的文献类型枚举值各不相同，需要指定 WOS、Scopus 枚举值到 NSTL 文献类型的映射方式。

如果 NSTL 字段可重复，外部数据源字段不可重复，直接根据映射字段取值即可。如果 NSTL 字段不可重复，外部数据源字段可重复，则需要指定解析规则从外部数据源中多个值中选择一个作为 NSTL 字段唯一值。例如将 Scopus 中的可重复字段 citation-language xml: lang=“”映射到 NSTL 中的不可重复字段 language，可设定为取第一个 citation-language 语种字段值。

从表 2 还可以看出，NSTL 通过元素方式进行描述，WOS、Scopus 多用属性进行描述，例如题名、页码、参考文献数都采用了属性限定元素的方式，更好地对描述内容进行归并。此外，WOS、Scopus 中期刊论文都具有唯一标识符，WOS 使用 uid 元素唯一标识论文，Scopus 使用 eid、pui、pii 等唯一标识论文。

表 2　论文元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
记录号	paper_id(R)	---	---
题名	title (R)	title type="item"	titletext original="y"
其他语种题名	alternative	title type="foreign"	titletext original="n"
文摘	abstract	abstract_text	abstract original="y"
其他语种文摘	abstract_alternative		abstract original="n"
关键词	keyword	keyword	author-keyword
其他语种关键词	keyword_alternative		
主题词	subject_heading	subject	mainterm
主题词表	thesaurus		descriptors controlled="y" type=""
分类号	classification		classification
分类法	classification_scheme		classifications type=""
正文语种	language (R)	language	citation-language xml: lang=""
其他语种	other_language		
起页	start_page (R)	page begin=""	pagerange first=""
止页	end_page	page end=""	pagerange last=""
总页数	total_page_number (R)	page page_count=""	pagecount
参考文献总数	total_reference_number	refs count=""	refcount=""
文献号	paper_no		
本地唯一标识符	local_doi (R)	---	---
DOI	doi	identifier type="doi" value=""	doi
论文类型	paper_type		
资源类型	type(R)	doctype	citation-type code=""

通过对 WOS、Scopus 元数据描述特点分析，对 NSTL Schema 进行局部修订，以属性限定元素的方式添加外部数据源论文唯一标识字段，与外部数据源此字段形成映射，例如添加 extend_ids extend_id type="" value=""，通过 type 属性与外部数据源唯一标识映射，value 为外部数据源唯一标识取值，一方面可以唯一识别来自于外部数据源的论文，与自加工数据进行区分，另一方面还为陆续添加其他数据源的唯一标识提供拓展框架。

3.2　作者/机构元数据的映射比较

在 NSTL 中，作者是指期刊论文撰写者，在 WOS、Scopus 中，论文作者与出版者、图表制作者、翻译者等共用子元素，因此需要指定角色类型或父元素才能实现准确映射，如表 3 所示。除了映射元素外，WOS、Scopus 中都有对作者姓、名、通讯作者、机构地址、所属国家和城市的描述，以及对作者唯一标识符如 ResearcherID、ORCID、AuthorID 等的描述。作者唯一标识符对唯一识别作者具有重要作用，可参考对论文唯一标识的处理方式，为 NSTL 添加外部数据源的作者唯一标识提供拓展框架。

从表 3 可以看出，在 6 个 NSTL 作者/机构元数据字段中，WOS、Scopus 均有 5 个字段可以映射，映射度较高。也存在同一字段在不同数据库取值类型不同的情况，例如在作者顺序字段，NSTL 的 author_sequence 取值类型为 byte[8]，WOS 中 seq_no 取值类型为 positive Integer，需要协调为一致才能真正地获取数据。

表 3　作者/机构元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
作者顺序	author_sequence (R)	name seq_no="" 且 role="author"	author seq=""
作者姓名	author_name(R)	name(role="author" addr_no="") display_name	author indexed-name
其他形式作者姓名	author_name_alternative	有 full_name 时对照 name(role="author") wos_standard; 无 full_name，无对照字段	author initials
作者所属机构	affiliation	address_name　address_spec (addr_no="") organization	affiliation organization
其他形式机构	affiliation_alternative		
作者 Email 地址	email	name(role="author") email_addr	author e-address type="email"

此外，在 NSTL 中顺序描述作者和机构信息，在 Scopus 中以机构为基准对作者进行划分，在 WOS 数据库中通过 addr_no 属性建立作者和机构之间的一一对应关系。如果作者姓名(name)元素中的属性 addr_no 和地址(address_spec)元素中的属性 addr_no 相同，则表示此机构是该作者的机构。这样，不管作者有几个机构，都可以方便地实现对应，避免重复记录。

3.3　期刊元数据的映射比较

期刊是期刊论文的载体，在 NSTL 中期刊元数据包括期刊描述元素(见表 4 中前 14 个字段)和卷期描述元素(见表 4 中后 3 个字段)，在 WOS、Scopus 中，卷期描述元素包含在期刊描述元素中。除了表 4 中的映射字段，WOS 包含了更详细的期刊名称缩写、卷期出版日期和出版商地址信息，Scopus 还描述了期刊唯一标识符 srcid、文献来源网址、期刊编辑者信息等。期刊、卷期唯一标识符对唯一识别期刊、卷期具有重要作用，可同样参考论文唯一标识的处理方式，为 NSTL 添加外部数据源的期刊、卷期唯一标识提供拓展框架。

表 4　期刊元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
母体数据源编号	catalog_code(R)	---	---
订购号	subscription_number		
ISSN	issn	identifier type="issn"	issn type="print"
EISSN	eissn	identifier type="eissn"	issn type="electronic"
CODEN	coden		codencode
国内统一书刊号	cn	identifier type="cn"	
母体文献名称	host_title(R)	title type="source"	sourcetitle
其他语种母体文献名称	host_title_alternative		translated-sourcetitle
语种	host_language(R)	---	---
母体文献分类号	host_classification		
出版地	publishing_place	publisher address_spec city	publisher affliation city
出版者	publisher	publisher name (role="publisher")display_name	publishername
起始年	start_year(R)	---	---
终止年	end_year		
卷期出版年	year(R)	pub_info pubyear=""	publicationyear first=""
卷信息	volume	pub_info vol=""	voliss volume=""
期信息	issue	pub_info issue="" part_no="" supplement="" special_issue=""	voliss issue="" supplement

在 NSTL 期刊元数据 17 个字段中，WOS 有 9 个字段实现映射，Scopus 有 10 个字段实现映射，对于未映射的必备字段处理方式同论文元数据未映射的必备字段处理方式。NSTL 中的一个元素可能对应于 WOS、Scopus 中的多个元素或同一元素中的多个属性。例如在 NSTL 中，只有期信息 issue 字段，没有划分增刊、

特刊、分期字段，但指定了这些字段在期信息字段中的著录规则，如有期号，但该期又分为若干分期的，分期前缀照录，增刊、专刊填写在期号后，若无期号则直接填写增刊信息等[9]，可根据这些著录规则对 WOS、Scopus 相应数据进行数据抽取合并。

3.4　参考文献元数据的映射比较

在 NSTL 中，参考文献内容包括引文作者、题名、出处、卷期以及获取访问路径等。参考文献信息可以让用户从作者研究脉络角度查找到一组相关文献[10]。WOS 包含了参考文献中的作者、题名、刊名、卷、页信息，没有参考文献原始信息字段，Scopus 既包含了原始信息字段，也包含了作者、题名等拆分字段。三者参考文献元数据映射如表 5 所示，对于未实现映射的必备字段处理方式同前。

表 5　参考文献元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
引文类型	citation_type(R)	---	---
引文原始信息	citation_orig_info(R)	---	ref-fulltext
引文第一作者	citation_author1	citedAuthor	ref-authors author seq=“1”
引文第二作者	citation_author2		ref-authors author seq=“2”
引文第三作者	citation_author3		ref-authors author seq=“3”
引文题名	citation_title	citedTitle	ref-titletext
引文出处	citation_sourcetitle	citedWork	ref-sourcetitle
引文出版年	citation_year	reference year	ref-publicationyear first=“”
引文卷号	citation_volume	reference volume	ref-volisspag voliss volume=“”
引文期号	citation_issue		ref-volisspag voliss issue=“”
引文页	citation_page	reference page	ref-volisspag pagerange first=“” last=“”
引文主编	citation_editor_in_chief		
引文出版者	citation_publisher		
链接地址	citation_url		

4　元数据映射方式的优势和不足

通过对 WOS、Scopus 与 NSTL 元数据的映射，可以看出，在大部分元数据字段上 WOS、Scopus 可以实现与 NSTL 元数据的映射，而且这些字段属于比较重要的字段。总体来说，通过元数据映射可以实现外部数据源数据到 NSTL 数据的准确转换，而且效率较高。因此，元数据映射是实现 NSTL 集成利用第三方来源元数据的可行方式和有效方式。元数据字段映射的数量越多，外部数据源数据利用越充分。

通过元数据映射的方式，还可以了解到其他数据库元数据字段的描述方式，与自有元数据规范进行比较，取长补短，提高自有元数据的完整性和兼容性。通过对 WOS、Scopus 元数据的分析和与 NSTL 元数据的映射，针对性地对现有 NSTL 元数据 Schema 进行修订。例如增加外部数据源数据唯一标识、修改元数据取值类型等，可以快速、高效、低成本地将外部数据源如 WOS 数据集成到 NSTL 联合数据加工系统中，也为后续增加其他第三方资源构建了拓展框架。

元数据映射方式虽然解决了三者数据库在信息组织方式和内容揭示方式上的部分差异，但依然存在局限性，例如，无法避免未能实现全部字段映射造成的目标信息丢失问题，会影响 NSTL 加工数据的全面性和完整性。又如，元数据描述的详略差异造成的源信息丢失问题，WOS、Scopus 对作者、机构、期刊等有更多较为详细的描述字段，在 NSTL 中没有体现，这些字段对文献资源的揭示更加细颗粒化，通过元数据映射输出的方式，造成外部数据源数据的丢失。

5　结　语

在当前不同文献数据库元数据描述字段不尽相同的情况下，如果相互之间的元数据能够进行映射，对实现不同数据库之间的数据交互和流转具有重要意义，元数据字段映射数量越多，数据越能得到充分利

用。本文以 WOS、Scopus 与 NSTL 期刊论文元数据的映射为基础，描述了 NSTL 集成利用第三方来源元数据的流程和方法，并提出元数据映射及集成第三方元数据过程中需要注意的问题。目前，NSTL 已将购买的 WOS 数据应用于数据加工过程中，陆续还会增加对其他数据源的数据应用，这对提升数据加工的速度和系统的自动化水平大有裨益。

参考文献：

[1] 张建勇，曾燕. 文献数据库数据加工规范[M]. 北京：知识产权出版社，2009. (Zhang Jianyong, Zeng Yan. NSTL Literature Data Processing Specification [M]. Beijing: Intellectual Property Publishing House, 2009.)

[2] 宋琳琳，李海涛. 大型文献数字化项目元数据互操作调查与启示[J]. 中国图书馆学报，2012，38(5)：27-38. (Song Linlin, Li Haitao. Metadata Interoperability in Mass Digitization Project: A Survey and Suggestions [J]. Journal of Library Science in China, 2012, 38(5): 27-38.)

[3] 申晓娟，高红. 从元数据映射出发谈元数据互操作问题[J]. 国家图书馆学刊，2006(4)：51-55. (Shen Xiaojuan, Gao Hong.Proceed from Metadata Mapping to Discuss Metadata Interoperability Problem [J]. Journal of the National Library of China, 2006(4): 51-55.)

[4] 萨蕾. 元数据互操作研究[J]. 情报科学，2014，32(1)：36-40. (Sa Lei. Study in Metadata Interoperability [J]. Information Science, 2014, 32(1): 36-40.)

[5] Web of Science [EB/OL]. [2014-05-08]. http://www.webofknowledge.com/WOS.

[6] Scopus [EB/OL]. [2014-06-18]. https://www.scopus.com/.

[7] The Singapore Framework for Dublin Core Application Profiles [EB/OL]. [2015-05-08]. http://dublincore.org/documents/singapore-framework/.

[8] NSTL_journalarticle.xsd [EB/OL]. [2015-05-20]. http://spec.nstl.gov.cn/specification/namespace/NSTL_journalarticle.xsd.

[9] Issue [EB/OL]. [2015-05-22]. http://spec.nstl.gov.cn/specification/index.php?title=Issue.

[10] Journal Article Metadata Specification [EB/OL]. [2015-06-05]. http://spec.nstl.gov.cn/specification/index.php?oldid.

作者贡献声明：

于倩倩：分析三个数据库元数据规范，进行映射比较，撰写并修订论文；
张建勇：提出集成利用第三方数据的基本框架和实现方案，提出论文修改意见。

收稿日期：2015-07-27
收修改稿日期：2015-09-06

Practices of NSTL Integrating and Using Third-party Metadata

Yu Qianqian　Zhang Jianyong
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] To apply the third-party source metadata such as Web of Science metadata to NSTL joint data processing system. [**Context**] Based on NSTL Development Program, need to expand process metadata by oneself to acquire metadata in various ways such as buying third-party metadata. [**Methods**] Map Web of Science, Scopus Schema to NSTL Schema, analyze the characteristics of Web of Science metadata to revise NSTL Schema. Based on mapping results, export third-party metadata as NSTL Schema format and integrat it into NSTL joint data processing system. [**Results**] Integrate the third-party metadata into NSTL joint data processing system rapidly and efficiently. [**Conclusions**] The apply of Web of Science metadata in NSTL joint data processing system has improved the data processing speed. Revising existing NSTL Schema targeted constructs widen fremwork for adding other third-party metadata.

Keywords: Web of Science　Scopus　NSTL　Metadata mapping